Tutorial

Scale-Space Theory for Multiscale Geometric Image Analysis

Bart M. ter Haar Romeny, PhD
Utrecht University, the Netherlands
B.terHaarRomeny@isi.uu.nl

Introduction

Multiscale image analysis has gained firm ground in computer vision, image processing and models of biological vision. The approaches however have been characterised by a wide variety of techniques, many of them chosen ad hoc. Scale-space theory, as a relatively new field, has been established as a well founded, general and promising multiresolution technique for image structure analysis, both for 2D, 3D and time series.

The rather mathematical nature of many of the classical papers in this field has prevented wide acceptance so far. This tutorial will try to bridge that gap by giving a comprehensible and intuitive introduction to this field. We also try, as a mutual inspiration, to relate the computer vision modeling to biological vision modeling. The mathematical rigor is much relaxed for the purpose of giving the broad picture. In appendix A a number of references are given as a good starting point for further reading.

The multiscale nature of things

In **mathematics** objects have no scale. We are familiar with the notion of points, that really shrink to zero, lines with zero width. In mathematics are no metrical *units* involved, as in physics. Neighborhoods, like necessary in the definition of differential operators, are defined as taken into the limit to zero, so we can really speak of *local operators*.

In **physics** objects live on a *range* of scales. We need an instrument to do an observation (our eye, a camera) and it is the range that this instrument can see that we call the scale range. To expand the range of our eye we have a wide armamentarium of instruments available, like microscopes and telescopes. The scale range known to humankind spans about 50 decades, as is beautifully illustrated in the book (and movie) "Powers of Ten" [Morrison 1985]. The range one instrument can see is always necessarily bounded on two sides: the *inner scale* is the smallest detail seen by the smallest aperture (e.g. one CCD element of our digital camera, a cone or rod on our retina); the *outer scale* is the coarsest detail that can be discriminated, i.e. it is the whole image (field of view).
In physics dimensional units are essential: we express any measurement in these units, like meters, seconds, candelas, ampères etc. There is no such thing as a physical 'point'.

In mathematics the smallest distance between two points can be considered in the limit to zero, but in physics this reduces to the finite aperture separation distance (sampling distance). Therefore we may foresee serious problems with notions as differentiation, especially for high order (these problems are known as regularization problems), sub-pixel accuracy etc. As we will see, these problems are just elegantly solved by scale-space theory.

In **front-end vision** the apparatus (starting at the retina) is equipped just to extract multi-scale information. Psychophysically is has been shown that the threshold modulation depth for seeing blobs of different size is constant (within 5%) over more than two decades, so the visual system must be equipped with a large range of sampling apertures. There is abundant electrophysiological evidence that the receptive fields (RF's) in the retina come in a wide range of sizes[1] [Hubel '62, '79a, '88a].

In any image analysis there is a *task*: the notion of scale is often an essential part of the description of the task: "Do you want to see the leaves or the tree"?

Linear Scale-Space Theory - Physics of Observation

To compute any type of representation from the image data, information must be extracted using certain *operators* interacting with the data. Basic questions then are: What operators to use? Where to apply them? How should they be adapted to the task? How large should they be? We will derive the kernel from first principles (axioms) below.

These operators (or filters, kernels, apertures: different words for the same thing) come up in many tasks in signal analysis. We show that they are a necessary consequence of the physical process of *measuring* data. In this section we derive from some elementary axioms a complete family of such filters. As we will see, these filters come at a continuous range of sizes. This is the basis of scale-space theory.

If we start with taking a close look at the observation process, we run into some elementary questions:

- What do we mean with the 'structure of images' [Koenderink 1984]?
- What is an image anyway?
- How good should a measurement (observation) be?
- How accurately can we measure?
- How do we incorporate the notion of scale in the mathematics of observation?
- What are the best apertures to measure with?
- Does the visual system make *optimal* measurements?

Any physical observation is done through an aperture. By necessity this aperture has to be finite (would it be zero no photon would come through). We can modify the aperture

[1] It is not so that every receptor in the retina (rod or cone) has its own fiber in the optic nerve to further stages. In a human eye there are about 150.10^6 receptors and 10^6 optic nerve fibres. Receptive fields form the elementary 'apertures' on the retina: they consist of many cones (or rods) in a roughly circular area projecting to a single (ganglion) output cell, thus effective integrating the luminance over a finite area.

considerably by using instruments, but never make it zero width. This implies that we never can observe the physical reality in the outside world, but we can come close. We can speak of the (for us unobservable) infinite resolution of the outside world.

We consider here physical observations by an initial stage measuring device (also called front-end) like our retina or a camera, where no knowledge is involved yet, no preference for anything, and no nonlinearities of any kind. We call this type of observation *uncommitted*. Later we will relax this notion, among others by incorporating the notion of a model or make the process locally adaptive to the image content, but here, in the first stages of observation, *we know nothing*.
This notion will lead to the establishment of *linear* scale-space theory. It is a natural requirement for the first stage, but not for further stages, where extracted information, knowledge of model and/or task comes in etc. We then come into the important realm of *nonlinear* scale-space theory, which will be discussed in section 4.

Scale-space theory is the theory of kernels where the size ('scale') of the kernel is a free parameter. This multi-scale approach is actually the natural physical situation. A single constant-size aperture function may be sufficient in a controlled physical application (e.g. fixed measurement devices), but in the most general case *no a priori size* is determined. Control is needed over the scale. Scale-space theory comes around whenever we observe our system of study, and thus it is applied at feature detection, texture, optic flow, disparity, shape, etc.

Noise is always part of the observation. We cannot separate it from the data. It can only be extracted from the observation if we have a model of the observed entity or the noise, or both. Very often this is not considered explicitly. One e.g. often assumes the object comes from the 'blocks world' so it has straight contours, but often this is not known.

In the next paragraph we consider the aperture function as an operator: we will search for *constraints* to pin down the exact specification of this operator. For an unconstrained front-end there is a unique solution for the operator: the Gaussian kernel. If there is no preferred size of the operator, it is natural to consider them at all sizes, i.e. as a family, parameterized by the parameter for scale (one per dimension): the standard deviation (the 'width' or 'scale') of the Gaussian kernel.

The aperture function of an uncommitted front-end
To derive the aperture function we first establish the requirements and constraints appropriate for the physical situation at hand.

Let us study an example where things go wrong: if we look at an image through a square aperture, and make larger copies of the aperture to do the sampling at larger scales, we get the blocky effect seen in figure 1. [Koenderink 1984a] coined this *spurious resolution*, i.e. the emergence of details that were there not before. We seem to have created new detail such as lines and corners in the image. By looking at the result through your eyelashes, you blur the image and remove the extra detail again. As a general rule we want the information content only to decrease with larger aperture.